Ankita Bhamidimarri
2211CS010025
S3-03

Employee Data Analysis Report

1. Dataset Description

1.1 Source: HR_Analytics dataset (1,000 records).

1.2 **Columns:**

- EmpID (string) - Employee ID, a unique identifier for each employee.
- Age (integer) - Age of the employee.
- Department (string) - Department in which the employee works.
- JobRole (string) - Role or position in the job.
- MonthlyIncome (double) - Monthly income of the employee.
- Attrition (string) - Employee attrition status (e.g., Yes or No).
- PerformanceRating (double) - Performance rating of the employee.

1.3 Data Quality:

- Schema inferred correctly; all columns present as expected.
- No explicit null/missing-value handling in notebook; basic exploration and previews performed.
- Types are consistent with intended use (IDs categorical, incomes numeric, ratings numeric).

2. Operations Performed

2.1 Data Cleaning & Exploration

- Loaded CSV into Spark DataFrame; previewed first rows; printed schema.
- Distinct counts: employees and departments; listed department names.
- Wrote filtered dataset to disk for Age > 45.

2.2 Descriptive Analytics

- Departmental employee count (pie chart): R&D 961 (65.4%), Sales 446 (30.3%), HR 63 (4.3%).
- Salary & Age distribution analysis (histogram, boxplots): MonthlyIncome is right-skewed with outliers up to ~20k; ages span from 18 with a notable 45+ segment.

2.3 Relationship Analysis

- Visualized Attrition counts by Department.
- Visualized MonthlyIncome distribution (histogram and boxplot).
- Visualized average PerformanceRating by JobRole.
- Applied a business rule transformation: +$500 to MonthlyIncome for employees with Age > 45 (for scenario analysis); inspected samples.

3. Key Insights

3.1 Workforce Demographics

- Organization size: 1,470 employees.
- Age spans include young entrants (example at 18) and a notable segment over 45 (extracted subset analyzed/saved)

3.2 Departmental Insights

- Research & Development is the largest unit (961 employees), followed by Sales (446); Human Resources is comparatively small (63).
- Attrition visualization created to compare departmental patterns (no numeric rates computed in-notebook).

3.3 Salary Insights

- Highest average incomes by role are in leadership: Managers (~17182) and Research Directors (~16034).
- Individual departments show wide salary ranges, with R&D exhibiting the broadest span (1009 to 19999).
- Several individual contributor roles (e.g., Research Scientist, Laboratory Technician, Sales Representative) have substantially lower average incomes.

4. Recommendations

4.1 Talent Development & Retention

- Use the attrition-by-department view to prioritize targeted retention programs; compute attrition rates next to identify hotspots.
- For roles with lower pay bands and stable performance (e.g., Lab Technician, Research Scientist), strengthen career pathways and upskilling to improve engagement.

4.2 Compensation Strategy

- Review pay structure for lower-paying roles relative to contribution and market, especially in Sales Representative and entry-level R&D roles.
- Monitor pay compression among top roles (Managers, Research Directors) and ensure performance-based differentiation is preserved.

4.3 Workforce Allocation

- Given R&D's size, assess workload distribution, managerial span of control, and support staffing to avoid bottlenecks.
- Evaluate whether HR headcount adequately supports organizational scale and needs.

4.4 Geographic Operations

- Add location attributes (country, region, city/site)
  to enable geographic analyses (pay parity, local attrition, hiring funnels)
  in future cycles

4.5 Future Analytics Opportunities

- Compute and track attrition rates by department/role/tenure; perform cohort and survival analyses.
- Incorporate additional drivers (tenure, education, overtime, training hours, promotions) and build predictive models for attrition risk.